DOCUMENT RESUME

ED 222 528 TM 820 676

AUTHOR Kolstad, Andrew
TITLE An Introduction to Event History Analysis.
PUB DATE 20 Mar 82
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 20, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Literature Reviews; *Longitudinal Studies; *Probability; *Statistical Analysis; *Systems Approach
IDENTIFIERS Continuous Assessment; Event Analysis; *Stochastic Analysis

ABSTRACT

The theory of stochastic processes deals with systems that develop over time in accordance with probabilistic laws. The basic concepts involved in two types of continuous-time processes are the idea of a constant probability of occurrence in the point event process and the extensions necessary for the discrete state process. The required mathematical skills and technical literature in this are discussed. It is recommended that researchers responsible for collecting longitudinal data change their method from a reference point to an event history approach to item construction. The difference between the two approaches is illustrated with sample questions from two longitudinal studies sponsored by the National Center for Education Statistics. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222528

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. Kolstad

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

AN INTRODUCTION TO EVENT HISTORY ANALYSIS

by
Andrew Kolstad
National Center for Education Statistics
400 Maryland Avenue, S.W. - DMES
Washington, D.C. 20202

TM 820 678

This paper was prepared for presentation to the Special Interest Group in Longitudinal Research at the annual meetings of the American Educational Research Association, New York City, March 20, 1982.

As people with a special interest in longitudinal studies, we are very concerned about how to study development and change over time. Today I will introduce some developments in the statistics of stochastic processes that make them particularly suitable for use with longitudinal data in education and other social sciences.

Some of you may remember reading papers in graduate school about Markov chain models for occupational mobility, industrial mobility, or educational progression. In such a paper you would have seen, for example, a table with the number of students enrolled in various kinds of schools in one year cross-classified with those enrolled in the next year. You then would have seen some equations with an unfamiliar mathematical form. The equations would somehow be applied to the table and projections of the future distributions of students among schools would be displayed. Such papers were infrequent because the results were not impressive. The projections were clearly inaccurate afer a fairly short interval; the model applied to the system as a whole, with no way to compare subgroups; and the ideas about why the world would behave like a Markov chain were unconvincing. A Markov chain is one kind of stochastic process, but we were right not to take that kind of work seriously in the social sciences.

Over the past ten years, however, a number of new developments in statistics and in sociology have drastically changed this picture. One change has been from a stochastic process in which time progresses in discrete jumps, to a model with a continuous flow of time. This change improved the realism of stochastic process models somewhat, in that changes would now take place at times that don't coincide with the observation times. But the major change

toward realism occurred with the development of multivariate methods and estimation techniques that model a world in which different kinds of people can change and develop in different ways at different speeds.

Tuma, Hannan, and Groeneveld (1979) reported from their research on effects of public welfare programs on family stability an example of just how successful a continuous-time, multivariate stochastic process model can be, by comparing its results with results from the more well-known multiple regression approach. Using data from the Seattle/Denver Income Maintenance Experiment, they estimated a continuous-time, stochastic process model in which female rates of marriage, marital dissolution, and attrition from the study were multivariate functions of welfare support levels, normal income, prior AFDC enrollment, children, age, education, and wage rates. In order to compare the prediction errors with a multiple regression analysis, they looked at seven outcome measures, (being married, single, or lost to the study, being continously married or single, and the number of marriages and dissolutions) at one year and at two years after the start of the experiment, regressed on initial marital status and the same other causal variables used in their stochastic process model. Even though their dynamic model was constrained to fit the entire time period, while the regression equations were free to fit each time point and outcome variable separately, they were surprised to find that for 10 of the 14 outcomes their model explained more of the sample variation than did linear regression analysis (1979, p. 841). Unlike multiple regression, however, stochastic process models predict the time path of change and development.

This example provides evidence of the claim I am making here today: that the statistics of stochastic processes for the social sciences have

developed to the point that those of us who work with longitudinal data need to make a serious effort to become acquainted with them. The rest of my paper covers three topics that are intended to aid this acquaintance. First, I introduce a few of the basic ideas of stochastic processes; second, I point to the essential reference papers describing the new methods and give some advice on how to read them; and third, I make a plea for some changes in the way longitudinal data are often collected, in order to make them more suitable to the new statistics.

A FEW BASIC IDEAS OF STOCHASTIC PROCESSES

The theory of stochastic processes deals with systems that develop over time in accordance with probabilistic laws. There are two kinds of continous-time processes of interest: The point event process in which events happen but no changes in state occur (the name comes from its representation of events as points on a time line), and the discrete state process in which a unit can change from one to another among a set of categories (because the possible values are categorical, the set of states is discrete; were the units free to take on any value, the state space would be continuous).

The point event process. I begin with the point event process because it is simpler. This kind of process could describe, for example, telephone calls arriving at a switch board, persons arriving at a line in a bank, or traffic accidents on a local highway. For this model, I consider the calls, the people, and the accidents to be independently generated and interchangable.

The chance that more than one event happens in a small time interval can be represented by a general function, $0(\Delta t)$, representing the rate at which multiple events occur, where Δt is the time interval. By assumption, this

function gets small much more quickly than the time interval, Δt. This assumption means that if we could measure time accurately enough, events occur one at a time, because this term goes to zero as Δt goes to the limit.

If the calls or people or accidents have a constant probability of arriving, with no trend or periodicity, then the chance that the event happens exactly once in a time interval, Δt, has two parts: $r\,\Delta t + 0(\Delta t)$, where r is the constant rate of occurrence. The chance that no event occurs is one minus the other two possibilities: $1 - r\Delta t + 0(\Delta t)$. (The coefficient and sign of the $0(\Delta t)$ term do not matter, because whatever they are, the term vanishes later). By assuming that events happen randomly with respect to time, the chance that an event happens in the next time interval is independent of any occurrences in previous intervals. These assumptions define what is called a Poisson process.

While the probability of an event is assumed to be constant with respect to time, time intervals between events are not expected to be constant; long intervals are much less likely to occur than short intervals. The distribution of durations between events can be derived from the above assumptions in a few steps. Let T be a random variable representing the duration until the first event after the starting time, and let P(t) be the probability that T is greater than an arbitrary time point, t:

$$P(t) = Pr(T>t).$$

Then for $\Delta t > 0$

$$P(t + \Delta t) = Pr(T > t + \Delta t)$$

$$= Pr(T > t \text{ and no event in the interval } (t, t + \Delta t))$$

$$= Pr(T > t)\, Pr(\text{no event in } (t, t + \Delta t) \mid T > t)$$

The independence assumption means that this conditional probability is unaffected by the condition, $T > t$, which refers to what happened before t, and the probability of no event in the interval, $(t, t + \Delta t)$ is $1 - r\Delta t + 0(\Delta t)$. Therefore,

$$P(t + \Delta t) = P(t)(1 - r\Delta t + 0(\Delta t)).$$

The derivative of the duration probability with respect to time can now be written as

$$\frac{dP(t)}{dt} = \lim_{\Delta t \to 0} \frac{P(t + \Delta t) - P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t) - r\Delta t\, P(t) + 0(\Delta t)\, P(t) - P(t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{-r\Delta t\, P(t) + 0(\Delta t)\, P(t)}{\Delta t}$$

Since $0(\Delta t)$ goes to zero more quickly than Δt, the second term disappears in the limit, and Δt divides out of the first term. What is left is a differential equation,

$$\frac{dP(t)}{dt} = -rP(t)$$

that can be solved by first separating terms,

$$\frac{dP(t)}{P(t)} = -rdt$$

and then integrating both sides

$$\int \frac{dP(t)}{P(t)} = -r \int dt.$$

Using the facts that $\int dx = x$ and $\int \frac{du}{u} = \ln u$, this becomes

$$\ln P(t) = -rt.$$

Exponentiating both sides gives the exponential form (so typical of stochastic processes) for the likelihood of a given duration between events:

$$P(t) = e^{-rt}.$$

I presented this derivation to illustrate several points. First, a very simple model, in which the chance of an event occuring is constant over time, generated this exponential form. With more complex models, of course, the

equations get more complicated, but the exponential form is very common. Second, the kind of mathematical tools that you need are somewhat different from the ones you use with linear models. You need to be able to work with exponentials, logarithms, and some calculus. Third, the rate parameter, r, governs how rapidly events occur. As the rate increases, the time between events decreases. In fact, the mean duration between events is 1/r. Rates of occurrence are bounded by zero at the lower end (even extremely long durations between events have a positive rate) and by a value of ten or fifteen at the upper end (beyond these values the durations between events get too small to measure).

In trying to understand how these processes work, it is very useful to simulate some cases, or "realizations". With discrete-time models, each step proceeds in constant jumps, but in continuous-time models, the time between events is variable. From the relationship, $P(t) = e^{-rt}$, the time intervals can be derived. A randor number generator U, that is uniform in the (0,1) interval, can provide time intervals that are equally likely, so

$$P(t) = e^{-rt} = U$$

$$-rt = \ln U$$

$$t = (\ln U)/r.$$

A simulation is useful for generating "data" from a model to help understand how the model works. Figure 1 (from Cox and Miller, 1965, p. 7) shows two realizations of this simple point-event process with the constant rate, r equal to 2.

FIGURE 1

An example of programming a simulation of a related continuous process can be found in Knott (1981).

It is also important to be able to go in the other direction, from having data to obtaining an estimate of the rate parameter. A straightforward estimate in this case is the total number of events divided by the total time the process was observed. A general problem arises with real data, however, in that the last observable time period is necessarily truncated. In Figure 2, the period of observation lies between the two vertical lines extending below the time axis at 0 and T. The times at which events occur are marked, as before, with an x. But the existence

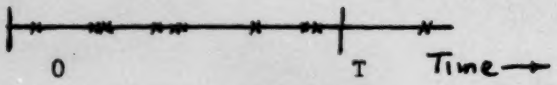

FIGURE 2

of the last event on the right is only assumed, since the event had not yet occurred at the observation time T. Whenever the last observed period does not end with an event, the period is said to be "right-censored." If the first period does not begin at the beginning of the process or with an event, the record is left-censored. Censoring is a characteristic of an observation plan, not the process itself. Sørensen (1977) and Tuma and Hannan (1978) discuss how to use data from censored time periods in such a way as to avoid bias in the estimators. Censoring is particularly important in social science research because longitudinal studies generally have short observation periods in relation to the rates of change in the outcomes of interest.

The point event model discussed so far needs to be extended for social science use in a way that permits different people to have different rates of change. This is done by letting the rate parameter, r, be a multivariate

function of background variables. In one of the earliest pieces of social science research of this kind*, Nancy Tuma (1976) modelled the rate of leaving a job as a multivariate linear function of job rewards and personal resources.**

Tuma (1976, pp. 342-3) discusses the conceptual advantage of modelling directly the rate of change parameter, rather than its observable consequences. While the transition rate is not itself observable, several observables can be derived from it: whether or not an event occurs by a given point in time, the duration of the between events, and the number of events in a given period of time. These three observables have each been modelled as a linear function of background variables. Tuma showed mathematically that the three types of variables cannot simultaneously be linear functions of background variables; if any one is a linear function, the other two can not be linear functions. Yet if the rate parameter is made a linear function of the background variables, values of the three observables can be derived simultaneously from a single model. The first goal of our study of stochastic processes should be to understand the nature and mathematical properties of the rate parameters.

---

* An earlier paper by Sørensen (1975) modelled job shifts as a function of single variable, race.

** A linear function has the undesirable property, however, of occasionally predicting negative rates of job mobility, which are mathematically undefined. To deal with this problem, Tuma later developed a method for constraining the rate parameter to positive values with a log-linear function:

$$\ln r = \sum_k b_k x_k$$

The discrete space process. While the discrete space process is more complex than the point event process, there are many similarities. This kind of process could describe, for example, a person changing from employed to unemployed to out of the labor force; a person changing from single to married to dropped out of the study; or a person changing from junior college to college to dropped out of school.

A discrete state process can model such changes in status. If Y(t) is a random variable whose value is the state occupied by a unit at time t, then for any two time points, u and t ($t < u$), and for any two states at these time points, one can define a transition probability, the likelihood that a unit will occupy state k at time u, given that it occupied state j at time t:

$$p_{jk}(t,u) = \Pr[Y(u) = k \mid Y(t) = j].$$

Since in this model time is continous, there is no advantage in arbitrarily choosing any two particular time points at which to examine this set of transition probabilities. The only way to handle such transitions in a general way is to rely on the assumption of constant rates of change, as in the point event processes, and show how the transition probabilities between any two points in time depend on these underlying rate parameters.

If one assumes that the probability of a transition occurring in a small time interval $(t, t + \Delta t)$ depends only on the state occupied at time t and not on any of the states occupied previously, then transition probabilities over long time intervals can be built up recursively from transition probabilities over shorter time intervals. Letting P(t,u) be a matrix whose elements are the $p_{jk}(t,u)$ above, this assumption means that for three time points,

$$P(s,u) = P(s,t)\, P(t,u), \quad s < t < u.$$

By letting the difference between t and u approach zero, this relationship can be used to define the derivative of the transition probabilities with respect to time:

$$\frac{dP(s,t)}{dt} = P(s,t)R,$$

where R is a matrix of constant rate parameters similar to that of the point event process. As before, the solution of this differential equation results in a equation with an exponential form,

$$P(s,t) = e^{R(t-s)},$$

that expressed the transition probabilities between any two time points as a function of the rate parameters and the elapsed time. *

Sometimes the analyst is also interested in the proportions of people (or units) occupying each state over time. Letting, the vector, P(t), represent the proportion of people in each state at time t, this relationship can be used to express the state probabilities as a function of the initial distribution, the rate parameters, and the elapsed time:

$$P(t) = P(0)\ P(0,t) = P(0)e^{Rt}.$$

A graph displaying an example of this function is contained in Tuma, Hannan, and Groeneveld (1979, p. 843), in which the authors plot the observed and the predicted proportions of married welfare mothers over a two year period.

---

* Raising the constant e to a matrix power is probably an unfamiliar operation for most people. In fact, a function of a square matrix can be expressed as the eigenvectors of the matrix times that function of its eigenvalues. More specifically, for the matrix A, if $A = B\ C\ B^{-1}$, where C is a diagonal matrix of eigenvalues and the columns of B are their corresponding eigenvectors, then $f(A) = Bf(C)B^{-1}$. In the case of this matrix exponential, this means

$$e^{Rt} = B\ \text{diag}\ (e^{c_1 t}, \ldots, e^{c_k t})B^{-1}.$$

The rate parameter of the point event process is in some respects similar to and in some ways different from the transition rates of the discrete space process. As before, the elements of the matrix of transition rates, r_{ij}, must be positive numbers greater than zero (and, in practical terms, less than about 10 or 15), with this exception: the diagonal elements are negative. The exponent in the equation for the point event process had a negative sign; in the discrete state process, this negative sign applies only to the diagonal elements, so it does not appear in the equation. The elements of the rate matrix are not all independent parameters. The diagonal elements are equal to, but opposite in sign from the sum of the other row elements:

$$-r_{jj} = \sum_{k} r_{jk}, \text{ for } k \neq j.$$

The duration of time spent in each state can differ from one state to another, but in each case the average duration is $1/r_{jj}$. In addition, the ratio r_{jk}/r_{jj} is the conditional probability that a change is to state k, given that a change out of state j occurs.

While the details of the above equations may not be well understood by those new to stochastic processes, it is important to understand that these transition rates are the essential quantities of interest in dealing with development and change that happens continously over time, because these quantities describe and govern the time path of changes. From the transition rates a number of observable characterstics can be derived exactly: the probability of occupying a given state at any given point in time, the expected duration in a state, and the number of changes in a given interval.

As in the case of the point event process, the discrete state process needs to be extended for use in the social sciences in a way that permits different kinds of people to develop in different ways. To introduce causal relationships, the dependence of the transition rates on the observable

variables can be specified in either of two ways. First, when the decision to leave the current state can be separated from the choice of destination, one can model the rate of leaving (the inverse of the duration) and the conditional destination rates:

$$\ln r_{jj} = \sum_h b_{jh} X_h$$

and

$$\ln r_{jk} = \sum_h c_{jh} X_h$$

Alternatively, if the rates of moving (the off-diagonal matrix elements) are not conceptually separate from the rates of staying (the diagonal matrix elements), a log-linear decomposition (to constrain the predicted rates to be positive) of all but the diagonal elements would be appropriate:

$$\ln r_{jk} = \sum_h b_{jkh} X_h, \text{ for } j \neq k.$$

It is because recent developments made this extension possible that we now can use this powerful class of statistical methods for longitudinal research in the social sciences.

I have a few general comments about estimation methods. First, unlike the old Markov chain models, the rate parameters are estimated not from grouped data in cross-tabulations, but from individual data from unit records. Second, maximum-likelihood or partial-likelihood techniques are used that have good properties even in small samples with a fair amount of censoring (Tuma and Hannan, 1978). These techniques make possible estimates of the standard errors for statistical tests of the signifigance of the coefficients expressing the dependence of the transition rates on the causal variables. Third, while the transition rates are unobservable, there need be no mystery to the estimation methods. As in the case of point events, the computations derive from the total number of transitions of each kind divided by the total amount of time a person was exposed to the possibility of such a transition. Finally,

computer programs are currently available, though not as a part of SPSS or SAS. Nancy Tuma 's program (1980) is available for a small fee from Nancy Tuma directly at Stanford University. James Coleman (1981) has also published programs for this kind of work. While I have not examined Coleman's programs in detail, my first impression is that Tuman's program is more friendly to the user.

LEARNING TO USE STOCHASTIC PROCESSES

Those of us who are persuaded that stochastic process models may be useful in their research and would like to learn to use them are faced with a difficult choice. In a way we are like potential buyers of a home microcomputer or an office word processor, asking ourselves whether we can afford to pass up immediate benefits to wait for the lower costs and improved performance that are likely a few years from now. Unless we need to use stochastic processes right now in our research (and members of this group are likely to have this need), there may be some advantage to waiting a few years while keeping an eye on further developments in this area. On the other hand, it took me longer to learn these methods than I expected, because the necessary mathematical skills are different from what I had been using, so I needed remedial work.

For those who want to start the process now, I offer this observation from my own experience. The first section of this paper illustrates my point that the mathematical skills required are more advanced than those required to use path analysis and structural equation models. One must be sufficiently exposed to elementary probability and statistics, matrix algebra, calculus, and differential equations to understand the basic concepts of stochastic processes. Unfortunately, there is as yet no introductory, graduate level

social science textbook for continuous-time, multivariate stochastic processes. Such a textbook in econometrics, for example, often contains a chapter on the basic results in matrix algebra that one needs to understand econometrics. The absence of such a well focussed textbook means learning some unneeded skills. In addition, the absence of such a textbook means we have to learn by reading a variety of technical journal articles that assume a generally higher level of mathematical sophistication than would be necessary for those who wish simply to apply these tecniques and not necessarily to contribute to their development.

I recommend that you begin your study with Hannan and Tuma's (1979) non-technical overview of methods for temporal analysis, and then proceed to some of the longitudinal research applications that have already begun to appear in the literature, in which the focus is on substance, with minimal attention on statistical technique. With these papers, you can set a minimal standard for what you need to learn. Here are five examples of longitudinal research applications: Hannan, Tuma, and Groenveld (1977) used this method to explain the effects of experimental negative income tax welfare programs on marital stability. Hannan and Carroll (1981) used this method to explain the effects of population, ethnic diversity, and gross national product on changes in the forms of government of the nations of the world. Sørensen and Tuma (1981) used this method to model the effects of ability, educational attainment, wages, and occupational standing on upward, downward, and lateral job changes. Rosenfeld (1981) used this method to model the ways in which career history, individual characteristics, and age affect transition rates between different types of jobs for men and women with advanced training in psychology. Felmlee (1982) used this method to model the effects of sex, job rewards, individual resources, social constraints, and age on rates of job changes within and between employers.

To acquire a knowledge of the formal assumptions, you need to study a textbook on stochastic processes, such as Cox and Miller (1965, Introduction and Ch. 4), Brieman (1973), Feller (1968, Ch. 17), Karlin and Taylor (1975), or perhaps Bartholomew (1973). One of these textbooks must be consulted, because none of the technical journal articles takes the space needed to explain how to get from the Markov assumption to the solution of the forward or backward Chapman-Kolmogorov differential equations.

To acquire an understanding of the properties of the matrix of transition intensities, the work of Singer and Spilerman (1974, 1976a, 1976b, 1978) is very useful. While these papers explain the advantages to changing from discrete-time Markov chains to continuous-time Markov processes and definitively treat the issues of embeddability and identification, this work is not practical for multivariate analysis because it treats only grouped data in cross-tabulations.

To learn how to analyze ungrouped longitudinal data with continuous-time stochastic processes, it makes sense to start with point event processes before moving to discrete state processes, as I did in the preceding section. Sørensen (1975) and Tuma (1976) discuss such models, and Sørensen (1977) and Tuma and Hannan (1978) explain what is to be done with censored data. The most important reference for discrete state models is the paper by Tuma, Hannan, and Groeneveld (1979), though it is probably too difficult for a beginner to follow. Tuman and Hannan should soon have a book out on the subject, and Coleman (1981) has recently published a book on continuous-time, discrete-state stochastic processes. While I have not yet finished the new Coleman book, so far I can report that it makes demands on mathematical skills that most beginners cannot meet. Issues of estimation are treated in Coleman (1981, Ch. 6), in Tuma, Hannan, and Groeneveld (1979), and in Tuma (1980).

For the past few summers, the National Opinion Research Center has been sponsoring, with some support from my office at NCES and the Labor Department, a summer short course in methods for analysis of longitudinal data, including the class of models I have discussed here. There are no other classes outside of the graduate schools of which I am aware.

COLLECTING LONGITUDINAL DATA

The next section of my paper is addressed to all those who collect longitudinal data, including those who want to learn these methods later, or perhaps leave this kind of analysis to the next generation of graduate students. This new class of methods requires slightly different questions in longitudinal surveys; it requires retrospective data on the timing of changes. The statistical estimates of the parameters of these models require data on the number of events of each type that occur to each person and the duration of time that each person was exposed to the possibility of each event. For example, these models need the dates people start and stop working at each job, the dates that people enter and leave each school they have attended, the dates of marriages and marital separations, and the dates of military service. If the intervals between survey waves are not long in comparison with the rates at which the events under study occur, questions of this type are not more space-consuming or burdensome than asking the respondents about their status at several different reference points. If the survey asks about each job, or each spell of schooling, then the analyst can assume there are no changes other than those listed by the respondents, and, consequently, one has full knowledge of the respondents' status at every point in time.

The use of stochastic process models in analysis requires a change in approach to collecting data, from asking about reference points to asking

complete histories. I can be more specific about this change by using as examples selected questionnaire items from NCES's two longitudinal studies, the National Longitudinal Study of the High School Class of 1972 (NLS-72) and the High School and Beyond (HS&B) study. The appendix to this paper contains selected education items from four follow-up surveys, the last three follow-ups from the NLS-72 and the first follow-up from HS&B.

The NLS-72 second follow-up needed to cover only a single year since the first follow-up, so it asked about school attendance in October 1974, a reference point, and other schools from October 1973 to October 1974, with beginning and ending dates of each. This was quite a reasonable and effective approach, provided that the respondent had not attended more than two schools in that period, and that the respondent did not drop out and then re-enter either school. In the latter case, the answer to "when did you first attend this school?" would not properly describe the episodes of schooling.

The NLS-72 third follow-up needed to cover two years since the second follow-up and, in order to ask about each reference point and to keep the questionnaire under a one-hour maximum required for EDAC clearance, the question about other schools at other times was deleted. So the respondent could report attending up to only two schools in this period, both of which had to be in October. This resulted in the net being too broad, especially for vocational students, for whom I estimatd that the average duration of schooling was about four-fifths of a year.*

---

* Using methods described in Singer and Spilerman (1976b), and using NLS-72 data to construct turnover tables between 4-year schools, 2-year schools, vocational schools and non-enrollment, for three pairs of the years 1972, 1973, and 1974, I extracted three continous-time transition intensity matrices. The inverse of the diagonal element for vocational school was about 0.8 years in each case. The results of this exercise were not published.

Another piece of evidence that suggests under-reporting of the flow of students through the postsecondary vocational sector is the unexpectedly high cummulative proportion of the NLS-72 cohort earning certificates and licenses (26.7 percent. See Kolstad, 1981). In addition, the date of leaving the school attended in October 1975 was not asked.

The NLS-72 fourth follow-up needed to cover three years since the third follow-up and the deficiencies of the reference point approach had become apparent. The item was changed to ask about school attendance for each month in each of the three years, but the school name was requested only for the last month attended, thus permitting good data on the October reference months. Unfortunately, the correspondence between this school and the prior months of attendance is temuous, especially when distinct spells of attendance were reported.

The HS&B first follow-up needed to cover two years since high school. This questionnaire abandoned all attempts to maintain October reference points, and asked about up to five different schools, even permitting simultaneous attendance in two schools.* When the HS&B first follow-up data become available in the next year we may begin to see event history analyses of student flows through the school system, because this is how the questions need to be asked in order to use stochastic process models of the kind I have discussed. I should point out that the National Longitudinal Survey of Young Americans, sponsored by the Labor Department, has also changed from a reference point to a retrospective history approach to labor market experience items.

---

* This may happen, for example, in training for the nursing profession; the student may be trained in a teaching hospital while taking academic courses in a junior college.

Once the data are collected in the form of a variable number of episodes, there are some choices to be made about how to structure the resulting data files. Blum, Karweit, and Sørensen (1969), Karweit (1973), and Ramsoy and Clarkson (1977) discuss the advantages of variable-length records for data storage, though standard packages like SPSS may have problems with this method.

I have argued in this section that longitudinal researchers, even if they do not themselves intend to learn and use stochastic process models, would be well-advised to collect their data in a retrospective event history form rather than a reference point form. It is time to discuss, briefly, what can be done with reference point data (sometimes known as panel data). This kind of data occurs in two situations: when the data were unfortunately collected the wrong way, or when the data could not be collected in the right way. The latter occurs when respondents are unaware of changes, or prior states are subject to gross recall errors. Some examples are prior attitudes, such as voting intentions or self-esteem, prior capacities, such as reading comprehension, or prior internal states, such as malaria parasites (Cohen and Singer, 1979). In these cases, the capacity for detailed multivariate analysis is much more limited. Coleman (1981, Ch. 4) addresses several topics in this area, using techniques for individual level data, while Coleman and Singer (1979) approach this problem from the more limited possibilities of group level data.

SUMMARY

In this paper I have argued that stochastic process models have recently advanced to the point that they have become useful for longitudinal research in the social sciences. I presented a few of the basic concepts involved in two types of continuous-time processes: the idea of a constant probability of occurance in the point event process and the extensions necessary for the discrete state process. In order to help those interested in learning more about this class of techniques, I presented some observations from my own experience on the required, level of mathematical skills and some guidance to the technical literature in this area. In order to permit more analysis of this kind, I asked those researchers responsible for collecting longitudinal data to change their methods from a reference point to an event history approach to item construction. I illustrated the difference between the two approaches with sample questions from two major longitudinal studies sponsored by the National Center for Education Statistics.

REFERENCES

Bartholomew, David J. 1973. "Stochastic Models for Social Processes." 2nd ed. New York: Wiley.

Blum, Zahava, Nancy Karweit, and Aage B. Sorenson. 1969. "A method for the collection and analysis of retrospective life histories" mimeo report No. 48, Center for the Social Organization of Schools, Johns Hopkins University.

Breiman, L. 1969. Probability and Stochastic Processes. Boston: Houghton Mifflin

Cohen, Joel, and Burton Singer. 1979. "Malaria in Nigeria: constrained continuous-time models for discrete-time longitudinal data on human mixed-species infections." Pp. 69-133 in S. Levin (ed.), Lectures on Mathematics in the Life Sciences Vol.12, Providence: American Mmathematical Society.

Coleman, James S. 1981. Longitudinal Data Analysis. New York: Basic Books

Cox, D. R., and H. D. Miller. 1965. The Theory of Stochastic Processes. New York: Wiley.

Feller, William. 1968. An Introduction to Probability Theory and Its Applications. Vol. 1. 3rd ed. New York: Wiley.

Felmlee, Diane, H. 1982. "Women's job mobility processes within and between employers." American Sociological Review 47 (February): 142-151.

Hannan, Michael T., and Glenn R. Carroll. 1981. "Dynamics of formal political structure: an event-history analysis." American Review of Sociology 46 (February): 19-35.

_____, and Nancy Brandon Tuma. 1979. "Methods for temporal analysis," Annual Review of Sociology 5:303-328.

_____, Nancy Brandon Tuma, and Lyle P. Groeneveld. 1977. "Income and marital events: evidence from an income-maintenance experiment." American Journal of Sociology 82 (May): 1186-1211.

Karlin, Samuel, and Howard Taylor. 1975. A First Course in Stochastic Processes. New York: Academic Press.

Karweit, Nancy. 1973. "Storage and retrieval of life history data." Social Science Research 2 (March): 41-50.

Knott, Gary D. 1981. "A study of a rotary queuing discipline." Interface: The Technical Notes Issued by the National Institutes of Health Computer Center, No 94, March, Pp. 36-38.

Kolstad, Andrew. 1981. "What college dropout and dropin rates tell us." American Education 17 (August/September): 31-33.

Ramsøy, Natalie Rogoff, and Stein-Erik Clausen. 1977. "Events as units of analysis in life history studies" paper prepared for the SSRC conference on the National Longitudinal Surveys, Washington, D.C.

Rosenfeld, Rachel A. 1981. "Academic men and women's career mobility." Social Science Research 10 (December): 55-75.

Singer, Burton, and Seymour Spilerman. 1974. "Social mobility models for heterogenous populations" Ch. 12 in H. L. Costner (ed.), Sociological Methodology, 1973-1974. San Francisco: Jossey-Bass.

_____. 1976a. "Some methodological issues in the analysis of longitudinal surveys." Annuals of Economic and Social Measurement 5 (Fall): 447-474.

_____. 1976b. "The representation of social processes by Markov models." American Journal of Sociology 82 (July): 1-54.

_____. 1978. "Clustering on the main diagonal in mobility matrices" Pp. 172-208 in K. F. Schuessler (ed.), Sociological Methodology 1979. San Francisco: Jossey-Bass.

Sørensen, Aage B. 1975. "The structure of intragenerational mobility." American Sociological Review 40 (August): 456-471.

_____. 1977. "Estimated rates from retrospective questions" Pp. 209-223 in D. R. Heise (ed.), Sociological Methodologoy 1977. San Francisco: Jossey-Bass.

_____, and Nancy Brandon Tuma. 1981 "Labor market structures and job mobility." Research in Social Stratification and Mobility: A Research Annual 1: 67-94.

Tuma, Nancy Brandon. 1976. "Reward, resources, and rates of mobility." American Sociological Review 41 (April): 338-360.

_____. 1980. "Invoking RATE". mimeo, Sociology Department, Stanford University

_____, and Aage B. Sørensen. 1978. "Approaches to the censoring problem in the analysis of event histories." Pp. 209-240 in K. F. Schuessler (ed.), Sociological Methodology 1979. San Francisco: Jossey-Bass.

_____, and Michael T. Hannan. forthcoming Social Dynamics: Models and Methods. New York: Academic Press

_____, Michael T. Hannan, and Lyle P. Groeneveld. 1979. "Dynamic analysis of event histories." American Journal of Sociology 84 (January): 820-854

APPENDIX

SCHOOL ATTENDANCE FROM OCTOBER 1973 THROUGH OCTOBER 1974

9. **From October 1973 through October 1974 were you enrolled in or did you take classes at any school like a college or university, service academy or school, business school, trade school, technical institute, vocational school, community college, and so forth?**

No.....1 *GO TO Q. 58, p. 10*
Yes .....2 *GO TO Q. 10* ⟶

10. **Did you attend school in the first week of October 1974?**

No .....1 *GO TO Q. 32, p. 7*
Yes.....2 *GO TO Q. 11*

11. **What is the exact name and location of the school you were attending in the first week of October 1974? (Please print and do not abbreviate.)**

School Name: _____

City: _____ State: _____

14. **When did you first attend this school?** _____ (month) _____ (year)

15. **Are you currently attending this school?**

Yes 1
No.....2 Date left: _____ (month) _____ (year)

ATTENDANCE AT OTHER SCHOOLS FROM OCTOBER 1973 TO OCTOBER 1974

32. **Besides any schools you may already have reported in this section, did you enroll in or take classes at any OTHER schools from October 1973 to October 1974? (Again include schools like colleges and universities, service academies, business schools, trade schools, technical institutes, vocational schools, community colleges, and so forth.)**

No.....1 *GO TO Q. 38, next page*
Yes .....2 *GO TO Q. 33*

33. **What is the exact name and location of this school? Please print and do not abbreviate. (If you attended more than one (other) school, then give the one that you attended the longest.)**

School Name: _____

City: _____ State: _____

35a. **When did you first attend this school?** _____ (month) _____ (year)

35b. **Are you now attending this school?**

Yes 1
No.....2 Date left: _____ (month) _____ (year)

51. **During the two-year period from October 1974 through October 1976 were you enrolled in or did you take classes at any school like a college or university, graduate or professional school, service academy or school, business school, trade school, technical institute, vocational school, community college, and so forth?**

No.....1 *GO TO Q. 98, p. 22*
Yes .....2 *CONTINUE WITH Q. 52*

SCHOOL ATTENDANCE IN OCTOBER 1976

52. **Did you attend school in the first week of October 1976?**

No.....1 *GO TO Q. 66, p. 15*
Yes .....2 *CONTINUE WITH Q. 53*

53. **What is the exact name and location of the school you were attending in the first week of October 1976? (Please print and do not abbreviate.)**

School Name: _____

City: _____ State: _____

55. **When did you first attend this school?** _____ (month) _____ (year)

56. **Are you currently attending this school?**

Yes 1
No.....2 Date left: _____ (month) _____ (year)

SCHOOL ATTENDANCE IN OCTOBER 1975

66. **Now please think back to Fall 1975. Were you taking classes or courses at any school during the month of October 1975?**

No .....1 *GO TO Q. 79, p. 17*
Yes, at the same school I attended in October 1976 and reported above in Q. 53.....2 *GO TO Q. 70*
Yes, at a school I have not yet reported .....3 *CONTINUE WITH Q. 67*

67. **What is the exact name and location of the school you were attending in October 1975? (Please print and do not abbreviate.)**

School Name: _____

City: _____ State: _____

69. **When did you first attend this school?** _____ (month) _____ (year)

SCHOOL ATTENDANCE DURING THE PERIOD FROM THE FIRST OF NOVEMBER 1976 THROUGH OCTOBER 1979

78. During the three-year period from the first of November 1976 through October 1979, were you enrolled in or did you take classes at any school such as a college or university, graduate or professional school, service academy or school, business school, trade school, technical institute, vocational school, community college, and so forth?

(Circle one.)

No.....1 *GO TO Q. 134, p. 30*
Yes.....2 *CONTINUE WITH Q. 79, p. 19*

SCHOOL ATTENDANCE DURING THE PERIOD FROM THE FIRST OF NOVEMBER 1978 THROUGH OCTOBER 1979

79. During the period from the first of November 1978 through October 1979, were you enrolled in or did you take classes at any school such as a college or university, graduate or professional school, service academy or school, business school, trade school, community college, and so forth?

(Circle one.)

No.....1 *GO TO Q. 91, p. 21*
Yes.....2 *CONTINUE WITH Q. 80*

80. During the period from the first of November 1978 through October 1979, which month(s) did you attend school?

(Circle all that apply.)

November 1978..... 1
December 1978..... 2
January 1979..... 3
February 1979..... 4
March 1979..... 5
April 1979..... 6
May 1979..... 7
June 1979..... 8
July 1979..... 9
August 1979.....10
September 1979.....11
October 1979.....12

81. What is the exact name and location of the school you attended the last month that you circled in Q. 80?

School name: _____

City: _____ State: _____

NOTE: Two similar blocks of items were asked for school attendance during the periods November 1977 to October 1978 and November 1976 to October 1977.

31. Between the time you left high school and the end of February 1982, have you enrolled in or did you take classes at any school such as college or university, graduate or professional school, service academy or school, business school, trade school, technical institute, vocational school, community college, and so forth? (Do not include Armed Forces training programs.) (MARK ONE)

Yes ..... ○ (GO TO Q. 32)
No ..... ○ (SKIP TO Q. 50)

32. Which months were you enrolled in or taking classes in any school between the time you left high school and the end of February 1982? (MARK ALL THAT APPLY)

1980	1981		1982
June ○	January ○	July ○	January ○
July ○	February ○	August ○	February ○
August ○	March ○	September ○	
September ○	April ○	October ○	
October ○	May ○	November ○	
November ○	June ○	December ○	
December ○			

33. Next we would like information about all of the schools you have gone to since you left high school. Please start with the first school you went to after high school. Answer questions A-K for that school in the first column (pages 16 and 18), then answer questions A-K for the second school in the next column, and so on. (BE SURE TO INCLUDE YOUR CURRENT SCHOOL.)

If you attended two schools at the same time, please put them in separate columns.

33. Continued.	COLUMN 1 1ST SCHOOL AFTER HIGH SCHOOL	COLUMN 2 2ND SCHOOL AFTER HIGH SCHOOL
A) What is the exact NAME and LOCATION of the school? (WRITE IN)	School name: _____ Address: _____ City: _____ State: _____	School name: _____ Address: _____ City: _____ State: _____
C) When did you START attending this school? (MARK OVALS FOR MONTH and YEAR)	Month ○ Jan. ○ May ○ Sept. ○ Feb. ○ June ○ Oct. ○ March ○ July ○ Nov. ○ April ○ Aug. ○ Dec. Year ○ 1980 ○ 1981 ○ 1982	Month ○ Jan. ○ May ○ Sept. ○ Feb. ○ June ○ Oct. ○ March ○ July ○ Nov. ○ April ○ Aug. ○ Dec. Year ○ 1980 ○ 1981 ○ 1982
D) When did you LEAVE this school? (MARK OVALS FOR MONTH and YEAR)	Am still attending this school, have NOT left ..... ○ Left in: Month ○ Jan. ○ May ○ Sept. ○ Feb. ○ June ○ Oct. ○ March ○ July ○ Nov. ○ April ○ Aug. ○ Dec. Year ○ 1980 ○ 1981 ○ 1982	Am still attending this school, have NOT left ..... Left in: Month ○ Jan. ○ May ○ Sept. ○ Feb. ○ June ○ Oct. ○ March ○ July ○ Nov. ○ April ○ Aug. ○ Dec. Year ○ 1980 ○ 1981 ○ 1982